Package 'ClusteredMutations'

April 29, 2016

Version 1.0.1

Date 2016-04-28

Title Location and Visualization of Clustered Somatic Mutations

Depends seriation

Description Identification and visualization of groups of closely spaced mutations in the DNA sequence of cancer genome. The extremely mutated zones are searched in the symmetric dissimilarity matrix using the anti-Robinson matrix properties. Different data sets are obtained to describe and plot the clustered mutations information.

License GPL-3

Author David Lora [aut, cre]

Maintainer David Lora <david@h12o.es>

NeedsCompilation no

Repository CRAN

Date/Publication 2016-04-29 07:44:36

R topics documented:

`dissmutmatrix` *Symmetric dissimilarity mutation matrix.*

Description

This function computes and returns the Euclidean distance matrix, where each cell represents the distance in base pairs between the chromosomal position of somatic mutations. The matrix can be used to graph the anti-Robinson matrix using the seriation technique (Hahsler and Hornik 2011). Plotting the distance matrix helps to visualize and identify mutation clusters in addition to locating the micro-clustered mutated regions within the macro-clustered mutated zones that occur during the oncogenic process.

Usage

```
dissmutmatrix(data = NULL, chr = NULL, position, subset = NULL, upper = FALSE)
```

Arguments

`data`	: somatic substitution mutations of the cancer genome data set.
`chr`	: chromosome where the somatic mutation is located.
`position`	: position of somatic mutations in the DNA sequence of the cancer genome.
`subset`	: chromosome where the distance between all somatic mutations will be calculated.
`upper`	: logical value indicating whether the upper triangle of the distance matrix should be printed.

Details

Color, in the posterior dissimilarity plot, is selected to visually identify hyper-mutated zones (min = 6; max=5000).

Value

`dissmutmatrix()` returns a symmetric dissimilarity matrix in base 10.

Author(s)

David Lora

References

Hahsler M, Hornik K and Buchta C (2008), Getting things in order: An introduction to the R package seriation. Journal of Statistical Software 25/3. Software available at http://www.jstatsoft.org/v25/i03/.

Hahsler M and Hornik K. Dissimilarity plots: A visual exploration tool for partitional clustering. Journal of Computational and Graphical Statistics, 10(2):335–354, June 2011.

See Also

dist, dissplot

Examples

```
require(seriation)

###Example 1:
example1<-c(1,101,201,299,301,306,307,317,318,320,418,518,528,628)
10**(dissmutmatrix(position=example1,upper=TRUE))
mut.matrix <- dissmutmatrix(position=example1)
dissplot(mut.matrix,method=NA,
 options=list(col = c("white","white","orange","orange","red","red","red")))


###Example 2:
###One hypermutated zone with Two hypermutated areas sharing somatic mutations.
example2<-c(1,101,201,299,301,306,307,317,318,320,402,404,406,628)
10**(dissmutmatrix(position=example2,upper=TRUE))
mut.matrix <- dissmutmatrix(position=example2)
dissplot(mut.matrix,method=NA,
 options=list(col = c("white","white","orange","orange","red","red","red")))


###Example 3:
#data(PD4107a)

###Visualizes a dissimilarity mutation matrix using seriation and matrix shading
###using the method developed by Hahsler and Hornik (2011).
###Chromosome 1;
#mut.matrix <- dissmutmatrix(data=PD4107a,chr=Chr,position=Position,subset=1)
#dissplot(mut.matrix, method=NA, options=list( col =
# c("black","navy","blue","cyan","green","yellow","orange","red",
# "darkred","darkred","white")))

###Chromosome 6;
#mut.matrix <- dissmutmatrix(data=PD4107a,chr=Chr,position=Position,subset=6)
#dissplot(mut.matrix, method=NA, options=list( col =
# c("black","navy","blue","cyan","green","yellow","orange","red","darkred",
# "darkred","white")))

###Chromosome 12;
#mut.matrix <- dissmutmatrix(data=PD4107a,chr=Chr,position=Position,subset=12)
#dissplot(mut.matrix, method=NA, options=list( col =
# c("black","navy","blue","cyan","green","yellow","orange","red","darkred",
#   "darkred","white")))
```

`features` *Describe the features of mutation in the hyper-mutated zones.*

Description

Several features were observed in the hyper-mutated zones, for example, kataegis is the proposed name for the hyper-mutated zones with a cluster of C>T and/or C>G mutations that are substantially enriched at TpCpN trinucleotides, on the same DNA strand and that co-localize with large-scale genomic structural variation (Alexandrov et al. 2013; Nik-Zainal et al. 2012).

Usage

```
features(data = NULL, chr = NULL, position, refbase, mutantbase,
 min = 6, max = 5000)
```

Arguments

`data`	: somatic substitution mutations of the cancer genome data set.
`chr`	: chromosome where the somatic mutation is located.
`position`	: position of somatic mutations in the DNA sequence of the cancer genome.
`refbase`	: reference base in the chromosome.
`mutantbase`	: the mutant base in the chromosome.
`min`	: a number `min` or more consecutive mutations.
`max`	: a distance less than or equal to a number `max` of bp.

Details

By default, `features()` identifies the mutations in the hyper-mutated zones (min = 6; max=5000). Complex mutations (Roberts et al. 2012; Roberts et al. 2013) are those segments containing >= 2 consecutive mutations with a distance =< 100 bp.

Value

`features()` returns a data set with all mutations in the hyper-mutated zones. The data set contains five variables:

`clustered`	: number of cluster.
`chr`	: chromosome.
`position`	: the position of mutation in the chromosome.
`ref_base`	: reference base in the chromosome.
`mutant_base`	: the mutant base in the chromosome.

Author(s)

David Lora

References

Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415-21.

Nik-Zainal S, Alexandrov LB, Wedge DC, et al; Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012 May 25;149(5):979-93.

Roberts SA, Sterling J, Thompson C, et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell. 2012 May 25;46(4):424-35.

Roberts SA, Lawrence MS, Klimczak LJ, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013 Sep;45(9):970-6.

Examples

```
data(PD4107a)
###Show the features of mutations in the hypermutated zones;
features(data=PD4107a,chr=Chr,position=Position,refbase=Ref_base,
mutantbase=Mutant_base)

###Locate complex mutations in the hypermutated zones;
kataegis<-features(data=PD4107a,chr=Chr,position=Position,refbase=Ref_base,
mutantbase=Mutant_base)
complex <-features(data=PD4107a,chr=Chr,position=Position,
refbase=Ref_base,mutantbase=Mutant_base,min=2,max=10)
sub.complex <-subset(complex,select=c(clustered,chr,position))
data.hyperm  <-merge(kataegis,sub.complex,by=c("chr","position"),all.x=TRUE)

###Summary of the number of complex mutations in the hypermutated zones;
table(data.hyperm$clustered.x)
table(data.hyperm$clustered.y)
table(data.hyperm$clustered.y,data.hyperm$clustered.x)
data.hyperm<-transform(data.hyperm,clustered =
ifelse(is.na(clustered.y)==FALSE,1,0))
table(data.hyperm$clustered.x)
table(data.hyperm$clustered,data.hyperm$clustered.x)
###All hypermutated zones had more than 6 mutations (without complex mutations).
```

`imd` *Calculate of the inter-mutational distance (IMD).*

Description

The inter-mutational distance (IMD) is the distance between each somatic substitution and the substitution immediately prior (Nik-Zainal et al. 2012). The `imd()` is used to graph the rainfall plot (Nik-Zainal et al. 2012). This graph localizes the regional clustering of mutations and represents the IMD on the y-axis on a log base 10 scale, with mutations ranked and ordered on the x-axis from the first variant on the short arm of chromosome 1 to the last variant on the long arm of chromosome X.

Usage

```
imd(data = NULL, chr = NULL, position, extra = NULL)
```

Arguments

data	: somatic substitution mutations of the cancer genome data set.
chr	: chromosome where the somatic mutation is located.
position	: position of somatic mutations in the DNA sequence of the cancer genome.
extra	: additional information to posterior graph.

Value

imd() returns a data set with six variables:

chr	: chromosome where the somatic mutation is located.
position	: position of somatic mutations in the DNA sequence of the cancer genome.
extra	: additional information.
number	: ordered mutations.
distance	: Euclidean distance between each somatic substitution and the substitution immediately prior (IMD) (Nik-Zainal et al. 2012).
log10distance	: Euclidean distance between each somatic substitution and the substitution immediately prior in base 10.

Author(s)

David Lora

References

Nik-Zainal S, Alexandrov LB, Wedge DC, et al; Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012 May 25;149(5):979-93.

Wickham H. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

See Also

plot

Examples

```
data(PD4107a)

###New variable
extra <- factor(c(),levels=c("T>C","T>G","T>A","C>T","C>G","C>A"))
extra[PD4107a$Ref_base=="A" & PD4107a$Mutant_base=="G"]<-"T>C"
extra[PD4107a$Ref_base=="T" & PD4107a$Mutant_base=="C"]<-"T>C"
```

```
extra[PD4107a$Ref_base=="A" & PD4107a$Mutant_base=="C"]<-"T>G"
extra[PD4107a$Ref_base=="T" & PD4107a$Mutant_base=="G"]<-"T>G"
extra[PD4107a$Ref_base=="A" & PD4107a$Mutant_base=="T"]<-"T>A"
extra[PD4107a$Ref_base=="T" & PD4107a$Mutant_base=="A"]<-"T>A"
extra[PD4107a$Ref_base=="G" & PD4107a$Mutant_base=="A"]<-"C>T"
extra[PD4107a$Ref_base=="C" & PD4107a$Mutant_base=="T"]<-"C>T"
extra[PD4107a$Ref_base=="G" & PD4107a$Mutant_base=="C"]<-"C>G"
extra[PD4107a$Ref_base=="C" & PD4107a$Mutant_base=="G"]<-"C>G"
extra[PD4107a$Ref_base=="G" & PD4107a$Mutant_base=="T"]<-"C>A"
extra[PD4107a$Ref_base=="C" & PD4107a$Mutant_base=="A"]<-"C>A"
PD4107a$extra<-extra

###Generate new data set with intermutational distance;
rainfall<-imd(data=PD4107a,chr=Chr,position=Position,extra=extra)

###Rainfall plot for PD4107a cancer sample;
plot(rainfall$number, rainfall$log10distance,pch=20,
ylab="Intermutation distance (bp)",xlab="PD4107a",yaxt="n",
col=c(rep(c("black","red"),14)[rainfall$chr]))
axis(2, at=c(0,1,2,3,4,6), labels=c("1","10","100","1000","10000","1000000"),
las=2, cex.axis=0.6)

###Rainfall plot for PD4107a cancer sample (Nik-Zainal et al. 2012);
#require(ggplot2)
#graph <- qplot(data=rainfall,number,log10distance,colour=extra, ylim=c(0.2,8),
# ylab="log10", xlab="PD4107a")
#graph <- graph +
# scale_colour_manual(values =c("T>C"="yellow","T>G"="green","T>A"="pink",
# "C>T"="red","C>G"="black","C>A"="blue"))
#graph <- graph + theme(legend.title=element_blank())
#graph <- graph + scale_y_continuous(breaks = c(0, 1, 2, 3, 4, 6),
# labels=c("1","10","100","1000","10000","1000000"))
#graph
```

PD4107a *Somatic mutations data set from a primary breast cancer genome.*

Description

PD4107a is a data set of somatic substitution mutations from a primary breast cancer whole genome with a germline mutation in BRCA1 (Nik-Zainal et al. 2012). The data set contains five variables: sample name, chromosome where the somatic mutation is located, location of the somatic mutation, the reference base and the mutated base.

The complete set of somatic mutations from a patient with breast cancer (PD4107a) was provided by the Cancer Genome Project group at the Wellcome Trust Sanger Institute (Alexandrov et al. 2013). Mutations with Indel labels were deleted (only subs).

Usage

```
data(PD4107a)
```

Format

A data frame with 9879 observations on the following 5 variables.

`Sample_id` : PD4107a.

`Chr` : From chromosome 1 to chromosome X.

`Position` : Mutation locations on the chromosome.

`Ref_base` : The reference base in the mutation locations.

`Mutant_base` : The mutated base in the mutation locations.

Details

Patient PD4107a has been described throughout the scientific literature (Alexandrov et al 2013; Fischer et al 2013; Muino et al 2014; Nik-Zainal et al 2012; Roberts et al 2013).

Source

ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl/somatic_mutation_data/Breast/Breast_clean_somatic_mutations_for_signature_analysis.txt

References

Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415-21.

Hahsler M and Hornik K. Dissimilarity plots: A visual exploration tool for partitional clustering. Journal of Computational and Graphical Statistics, 10(2):335–354, June 2011.

Fischer A, Illingworth CJ, Campbell PJ, et al. EMu: probabilistic inference of mutational processes and their localization in the cancer genome. Genome Biol. 2013 Apr 29;14(4):R39.

Muino JM, Kuruoglu EE, Arndt PF. Evidence of a cancer type-specific distribution for consecutive somatic mutation distances. Comput Biol Chem. 2014 Aug 23. pii: S1476-9271(14)00091-7.

Nik-Zainal S, Alexandrov LB, Wedge DC, et al; Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012 May 25;149(5):979-93.

Roberts SA, Lawrence MS, Klimczak LJ, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013 Sep;45(9):970-6.

Examples

```
data(PD4107a)

###PD4107a data set;
head(PD4107a,12)

###generate new data set with intermutational distance;
#rainfall<-imd(data=PD4107a,chr=Chr,position=Position)
###Rainfall plot for PD4107a cancer sample;
#plot(rainfall$number, rainfall$log10distance,pch=20,
# ylab="Intermutation distance (bp)",xlab="PD4107a",yaxt="n",
```

```
# col=c(rep(c("black","red"),14)[rainfall$chr]))
#axis(2, at=c(0,1,2,3,4,6), labels=c("1","10","100","1000","10000","1000000"),
# las=2, cex.axis=0.6)

###Locate the clustered mutations;
#showers(data=PD4107a,chr=Chr,position=Position)

###Visualizes a dissimilarity mutation matrix using seriation and matrix shading
### using the method developed by Hahsler and Hornik (2011).
###Chromosome 6;
#mut.matrix <- dissmutmatrix(data=PD4107a,chr=Chr,position=Position,subset=6)
#dissplot(mut.matrix, method=NA, options=list( col =
#  c("black","navy","blue","cyan","green","yellow","orange","red",
# "darkred","darkred","white")))
```

---

showers *Location of clustered mutations in the cancer genome.*

Description

`showers()` identifies all groups of closely spaced mutations using the anti-Robinson matrix. Hypermutated regions are defined as those segments containing a number (min = 6) or more mutations with a distance that is less than or equal to a number (max=1000) of bp, and referred to as mutation showers (Drake 2007a; Wang et al. 2007), clustered mutations (Drake 2007a; Drake 2007b; Roberts et al. 2012), or kataegis (from the Greek word for shower or thunderstorm) (Alexandrov et al. 2013; Nik-Zainal et al. 2012). `showers()` can be used to locate complex mutations (Roberts et al. 2012; Roberts et al. 2013) (min = 2; max=10).

Usage

```
showers(data = NULL, chr = NULL, position, min = 6, max = 5000)
```

Arguments

`data`	: somatic substitution mutations of the cancer genome data set.
`chr`	: chromosome where the somatic mutation is located.
`position`	: position of somatic mutations in the DNA sequence of the cancer genome.
`min`	: a number `min` of consecutive mutations.
`max`	: a distance less than or equal to a number `max` of bp.

Details

By default, `showers()` identifies the hyper-mutated zones (min = 6; max=5000). Complex mutations are those segments containing >= 2 consecutive mutations with a distance =< 100 bp.

Value

`showers()` returns a data set with all hyper-mutated zones in the DNA sequence of tumor cells.

`chr`	: the shower mutations data set contains seven variables: chromosome.
`pend`	: the last position in the chromosome of the mutation shower.
`pstart`	: the first position in the chromosome of the mutation shower.
`nend`	: the last number of a consecutive mutation shower.
`nstart`	: the first number of a consecutive mutation shower.
`distance`	: the length of a hyper-mutated zone and the number of mutations in the clustered mutation.

Author(s)

David Lora.

References

Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. Nature. 2013 Aug 22;500(7463):415-21.

Drake JW. Mutations in clusters and showers. Proc Natl Acad Sci U S A. 2007 May 15;104(20):8203-4.

Drake JW. Too many mutants with multiple mutations. Crit Rev Biochem Mol Biol. 2007 Jul-Aug;42(4):247-58.

Nik-Zainal S, Alexandrov LB, Wedge DC, et al; Breast Cancer Working Group of the International Cancer Genome Consortium. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012 May 25;149(5):979-93.

Roberts SA, Sterling J, Thompson C, et al. Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions. Mol Cell. 2012 May 25;46(4):424-35.

Roberts SA, Lawrence MS, Klimczak LJ, et al. An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. Nat Genet. 2013 Sep;45(9):970-6.

Wang J, Gonzalez KD, Scaringe WA, et al. Evidence for mutation showers. Proc Natl Acad Sci U S A. 2007 May 15;104(20):8403-8.

Examples

```
###Example 1:
example1<-c(1,101,201,299,301,306,307,317,318,320,418,518,528,628)
showers(position=example1, min=5, max=100)

###Example 2:
example2<-c(1,101,201,299,301,306,307,317,318,320,402,404,406,628)
showers(position=example2, min=5, max=100)

###Example 3:
#data(PD4107a)
```

```
###Locate the clustered mutations;
#showers(data=PD4107a,chr=Chr,position=Position)

###Locate the complex mutations;
#complex.showers<-showers(data=PD4107a,chr=Chr,position=Position,min=2,max=10)
#nrow(complex.showers)
#table(complex.showers$chr)
```

Index